Choice Models in Operations

Lecture 4 : Random Utility Maximization (RUM) Models

Instructor: Srikanth Jagabathula *Scribe: Haotian Song*

A quick review: from last lecture, we have

$\mathbb{P}(\cdot|\cdot)$ (stochastic choice function)
$\uparrow$ (ARSP)
Rational (consistent with a distribution over preference lists)
$\updownarrow$ (finite product space)
Random Utility Representation ($\mathbb{P}(1 \succ 2 \succ \cdots \succ n) = \mathbb{P}_F(U_1 \geq U_2 \geq \cdots \geq U_n)$)

where we assume the distribution is continuous.

Now we focus on specific models. Specifying a choice model that is rational is equivalent to specifying a distribution over preference lists.

Example 1. To specify the choice function for choosing between coffee and tea, we need to specify the probability that the underlying preference list is

$$\sigma_1 : \text{coffee} \succ \text{tea} \quad \text{and} \quad \sigma_2 : \text{tea} \succ \text{coffee}.$$

Suppose we specify that $\mathbb{P}(\sigma_1) = 60\%$ and $\mathbb{P}(\sigma_1) = 40\%$. This implies that in each choice instance the customer samples σ_1 with probability 60% or σ_2 with probability 40%, and that

$$\mathbb{P}\{\text{coffee}|(\text{coffee,tea})\} = 60\% = \mathbb{P}(\sigma_1).$$

We can also equivalently specify the model in terms of utilities. For example, suppose the customer samples the utility vector

$$\begin{bmatrix} U_{\text{coffee}} \\ U_{\text{tea}} \end{bmatrix} = \begin{cases} \begin{bmatrix} 40 \\ 20 \end{bmatrix} & \text{with probability } 60\% \\ \begin{bmatrix} 10 \\ 30 \end{bmatrix} & \text{with probability } 40\% \end{cases}$$

which yields the same stochastic choice function.

Because a distribution over preference lists can be equivalently specified through a distribution over utilities, historically two approaches have been taken. In the **first** approach, distributions over rankings were directly specified.

Example 2. [**Distance-Based Models**] A popular way to specify a distribution over rankings is to use a *distance* (can be pseudo-distance) function over preference lists. In particular, suppose $d(\sigma_1, \sigma_2)$ specifies "*how far*" the two rankings are. For instance, the **Kendall-Tau Distance** measures the number of pairwise disagreements between two rankings:

$$d(\sigma_1, \sigma_2) = \sum_{1 \le i < j \le n} \mathbb{1}[(\sigma_1(i) - \sigma_1(j))(\sigma_2(i) - \sigma_2(j)) < 0],$$

where $\sigma(i)$ is the preference rank of item i under ranking σ.

In Example 1, letting i=coffee and j=tea, with lower rank prefered, we have

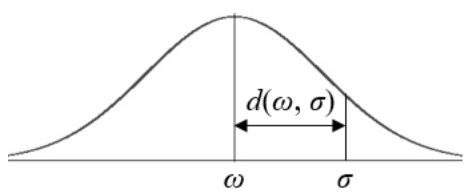
$$\begin{aligned} \sigma_1(i) - \sigma_1(j) &= 1 - 2 = -1 \\ \sigma_2(i) - \sigma_2(j) &= 2 - 1 = 1 \\ d(\sigma_1, \sigma_2) &= \mathbb{1}[(\sigma_1(i) - \sigma_1(j))(\sigma_2(i) - \sigma_2(j)) < 0] = 1 \end{aligned}$$

which implies there is pairwise disagreement between σ_1 and σ_2 on the preference of i and j.

Also note the maximum distance between two rankings of n items is $\binom{n}{2}$, such as

$$\begin{aligned} &\sigma_1 : 1 \succ 2 \succ \cdots \succ n \\ \text{and} \quad &\sigma_2 : n \succ n-1 \succ \cdots \succ 1 \end{aligned}$$

Given a distance function, we consider the following distribution $\mathbb{P}(\sigma) \propto e^{-\theta d(\sigma, \omega)}$ where $\theta > 0$ is the concentration parameter and $\omega \in \mathscr{S}_n$ is the modal ranking on the centroid ranking. This distribution is unimodal and has a mode at ω.



When the distance function is Kendall-Tau, this model is called the *Mallows Model* which is very popular in machine learning and statistics but not in operations and marketing. The concentration parameter $\theta = 0$ yields a uniform distribution and large values of θ result in concentrated or "*peaked*" distributions. The modal ranking ω makes the model hard to use in practice.

Example 3. If we are given m samples $\sigma_1, \sigma_2, \cdots, \sigma_m$, then it can be shown that

$$\omega \in \operatorname*{argmin}_{\omega \in \mathscr{S}_n} \sum_{i=1}^{m} d(\sigma_i, \omega)$$

when ω is estimated using maximum likelihood.

We write $\mathbb{P}(\sigma) = \frac{1}{z(\theta)} e^{-\theta d(\sigma,\omega)}$ where $z(\theta) = \sum_{\sigma \in \mathscr{S}_n} e^{-\theta d(\sigma,\omega)}$ is the normalized constant and is also called the partition function. By exploiting the symmetry of the distance function, it can be shown that $z(\theta)$ does not depend on ω. In particular,

$$z(\theta) = \prod_{i=1}^{n} \frac{1 - e^{-i\theta}}{1 - e^{-\theta}}.$$

[Reference: a monograph by *Persi Diaconis*]

We now take the **second** approach and specify the choice model by specifying a distribution over the utility vectors. This approach has been very popular first in psychology, transportation, economics and then in marketing and operations. This approach yields the **Random Utility Maximization (RUM)** class of models. Here we specify the utility of each of items and then a joint distribution over all the utilities. In particular, we say

$$U_i = \text{utility for product } i = V_i + \varepsilon_i$$

where we split the utility of item i into a deterministic component V_i and a random component ε_i which is also called the "*error*" term. Note that the decomposition is **not** unique.

Why people call them the error terms? Historically, RUM models were first used by *Thurstone* in 1920s. He was doing some sound experiments and in Thurstonian models all the errors are independent. In 1970s, the MNL models (see below) became popular and the error terms referred to something that the modeller could not observe.

Once the deterministic components are specified, we need to specify a joint distribution over the error terms. We start with the binary case with $n = 2$:

$$U_1 = V_1 + \varepsilon_1 \quad \text{and} \quad U_2 = V_2 + \varepsilon_2.$$

To compute choice probabilities, we could compute the probabilities of preference lists induced by the utility specification and then compute choice probabilities, **or** we could directly compute the choice probabilities. These two computations yield the same result when $n = 2$. So,

$$\begin{aligned}\mathbb{P}(1 \succ 2) = \mathbb{P}(1|\{1,2\}) &= \mathbb{P}(U_1 \geq U_2) \\ &= \mathbb{P}(V_1 + \varepsilon_1 \geq V_2 + \varepsilon_2) \\ &= \mathbb{P}(\varepsilon \leq V_1 - V_2) = F_\varepsilon(V_1 - V_2)\end{aligned}$$

where $\varepsilon = \varepsilon_2 - \varepsilon_1$ and F_ε is its CDF.

- If $\varepsilon \overset{D}{\sim} \mathcal{N}(0, \sigma^2)$, then we get the **Binary Probit Model**.
- If $\varepsilon \overset{D}{\sim} Logistic(0, \mu)$, then we get the **Binary Logit Model**.

The normal distribution does not have an explicit formula for its CDF, while the logistic distribution does:

$$F_\varepsilon(x) = \frac{1}{1 + e^{-\mu(x-\eta)}}, \quad \text{for } \varepsilon \overset{D}{\sim} Logistic(\eta, \mu)$$

where η is the location parameter and $\mu > 0$ is the scale parameter. In particular, substituting into the expression above, we obtain

$$\mathbb{P}(1 \succ 2) = \frac{1}{1 + e^{-\mu(V_1 - V_2)}} = \frac{e^{\mu V_1}}{e^{\mu V_1} + e^{\mu V_2}}.$$

For the binary case, we need not know the distributions of ε_1 and ε_2; we only need the distribution of ε. When there are n items, the binary probit and logit models generalize to the **Multinormial Probit Model (MNP)** and **Multinormial Logit Model (MNL)** respectively. In particular,

- an MNP model specifies that $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) \overset{D}{\sim} \mathcal{N}(0, \sum_{n \times n})$;
- an MNL model assumes that $(\varepsilon_i)_{i=1}^n$ are i.i.d. **Gumbel** distributed.

The Gumbel distribution is Type I extreme value distribution, the other two types being **Fréchet** (Type II) and **Weibull** (Type III).

These distributions arise as the limit distribution of appropriately randomized maximum of a sequence of i.i.d. random variables. In particular, suppose $X_1, X_2, \cdots, X_n$ are i.i.d. and $M_n = \max\limits_{1 \le i \le n} X_i$. If

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \le x\right) \to G(x) \quad \text{as } n \to \infty$$

for some sequences of constants $\{a_n\}$ and $\{b_n\}$, then $G(\cdot)$ must be one of the three extreme value distributions. In short,

- The **Central Limit Theorem (CLT):** taking the limit of the average of a sequence of random variables as $n \to \infty$, we obtain the normal distribution.
- The **Extreme Value Theory (EV):** taking the limit of the maximum of a sequunce of random variables as $n \to \infty$, we obtain one of the three types.

The EV theory was developed in 1950s.

For $\varepsilon \overset{D}{\sim} Gumbel(\eta, \mu)$, its CDF is

$$F_\varepsilon(x) = \exp(e^{-\mu(x-\eta)})$$

where η is the location parameter and $\mu > 0$ is the scale parameter.

Properties of the Gumbel distribution

1. The mode is η.
2. The mean is $\eta + \dfrac{\gamma}{\mu}$, where γ is the Euler's constant

$$\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \ln n\right) \approx 0.57721.$$

3. The variance is $\dfrac{\pi^2}{6\mu^2}$.
4. The affine transformation of a Gumbel is still Gumbel: if $\varepsilon \overset{D}{\sim} Gumbel(\eta, \mu)$, then

$$\alpha\varepsilon + V \overset{D}{\sim} Gumbel\Big(V + \alpha\eta, \frac{\mu}{\alpha}\Big),\ \forall\alpha > 0 \text{ and } V \in \mathbb{R}.$$

5. The difference of independent Gumbels with the same scale is logistic: if

$$\varepsilon_1 \overset{D}{\sim} Gumbel(\eta_1, \mu) \quad \text{and} \quad \varepsilon_2 \overset{D}{\sim} Gumbel(\eta_2, \mu)$$

are independent, then their difference

$$\varepsilon = \varepsilon_1 - \varepsilon_2 \overset{D}{\sim} Logistic(\eta_1 - \eta_2, \mu).$$

Note that there are also other distributions whose difference is logistic.

6. The maximum of independent and same-scale Gumbels is still Gumbel: if

$$\varepsilon_1 \overset{D}{\sim} Gumbel(\eta_1, \mu) \quad \text{and} \quad \varepsilon_2 \overset{D}{\sim} Gumbel(\eta_2, \mu)$$

are independent, then their maximum

$$\varepsilon = \max(\varepsilon_1, \varepsilon_2) \overset{D}{\sim} Gumbel\Big(\frac{1}{\mu}\ln(e^{\mu\eta_1} + e^{\mu\eta_2}), \mu\Big).$$

This is the most critical property which means Gumbel distributions are closed under maximization. Also recall that exponential distributions are closed under minimization. By induction, it follows that if $\varepsilon_i \overset{D}{\sim} Gumbel(\eta_i, \mu)$, $1 \le i \le n$ are independent, then

$$\varepsilon = \max_{1\le i\le n} \varepsilon_i \overset{D}{\sim} Gumbel\Big(\frac{1}{\mu}\ln\sum_{i=1}^{n} e^{\mu\eta_i}, \mu\Big).$$

Note that the mode is called the **log-sum-exp** function which is convex in η_i's.

From this, we derive the choice probabilites. Assume $\varepsilon_i \overset{i.i.d.}{\sum} Gumbel(0, \mu)$, $1 \le i \le n$. By

Property 4, we have

$$U_i = V_i + \varepsilon_i \overset{D}{\sim} Gumbel(V_i, \mu).$$

Then,

$$\begin{aligned}
\mathbb{P}(1|\{1, 2, \cdots, n\}) &= \mathbb{P}(U_1 \geq U = \max_{2 \leq i \leq n} U_i) \\
&= \mathbb{P}(U - U_1 \leq 0) \ (U \overset{D}{\sim} Gumbel\Big(\frac{1}{\mu} \ln \sum_{i=2}^{n} e^{\mu V_i}, \mu\Big) \text{ by Property 6}) \\
&= F(0) \ (F \text{ is the CDF of } Logistic\Big(\frac{1}{\mu} \ln \sum_{i=2}^{n} e^{\mu V_i} - V_1, \mu\Big) \text{ by Property 5}) \\
&= \frac{1}{1 + e^{\mu\Big(\frac{1}{\mu} \ln \sum_{i=2}^{n} e^{\mu V_i} - V_1\Big)}} \\
&= \frac{e^{\mu V_1}}{e^{\mu V_1} + e^{\mu V_2} + \cdots + e^{\mu V_n}}.
\end{aligned}$$